Applied Deep Learning

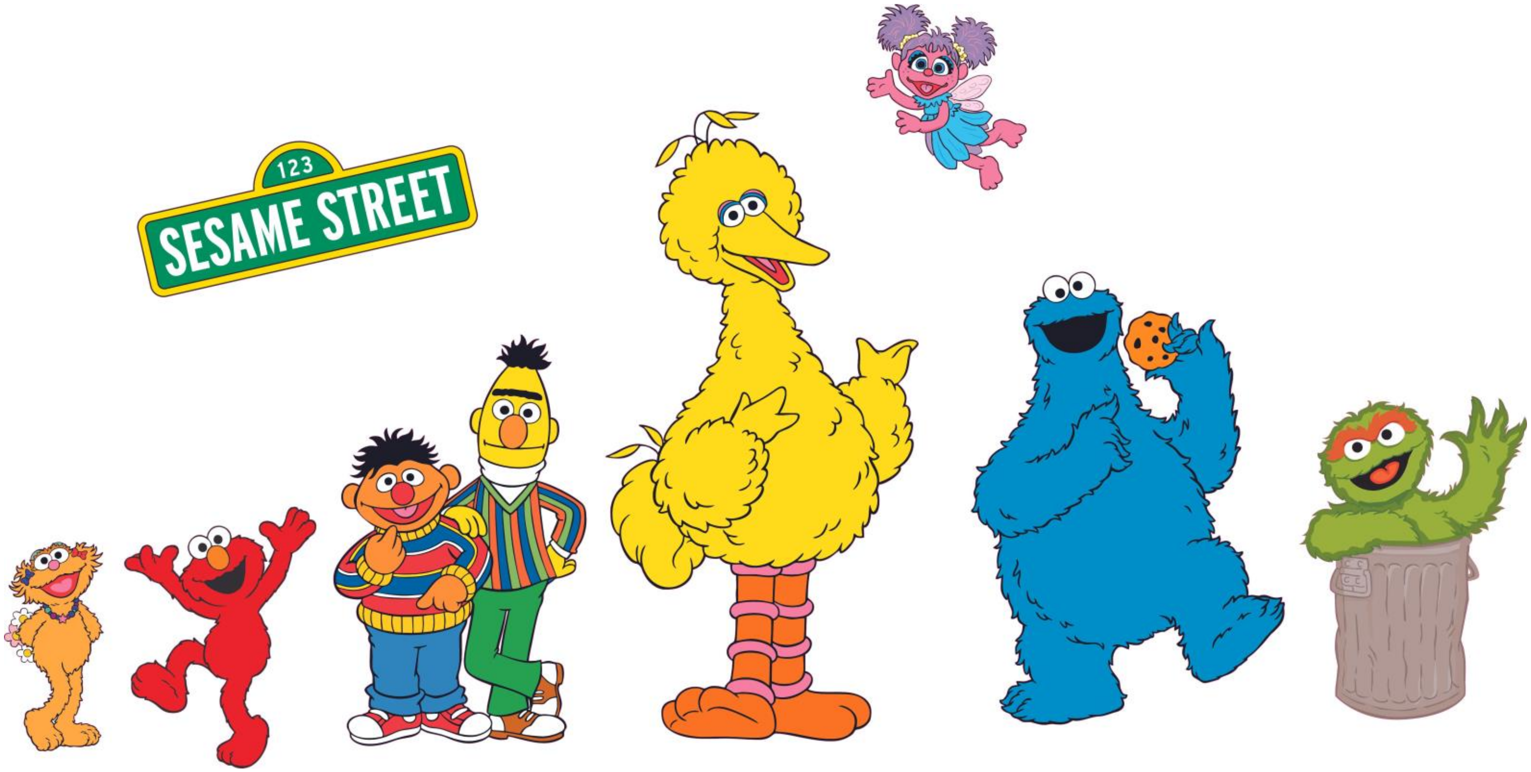
BERT

Bidirectional Encoder Representations from Transformers

October 12th, 2023 http://adl.miulab.tw

National Taiwan University
國立臺灣大學

Sesame Street



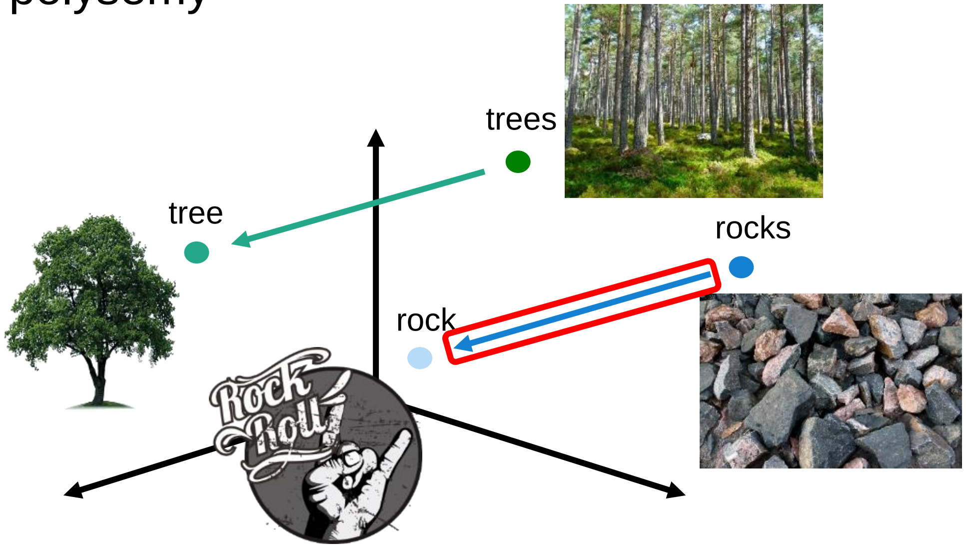
ELMO BERT

Picture from https://imagizer.imageshack.com/img924/8457/xhILHR.jpg

3

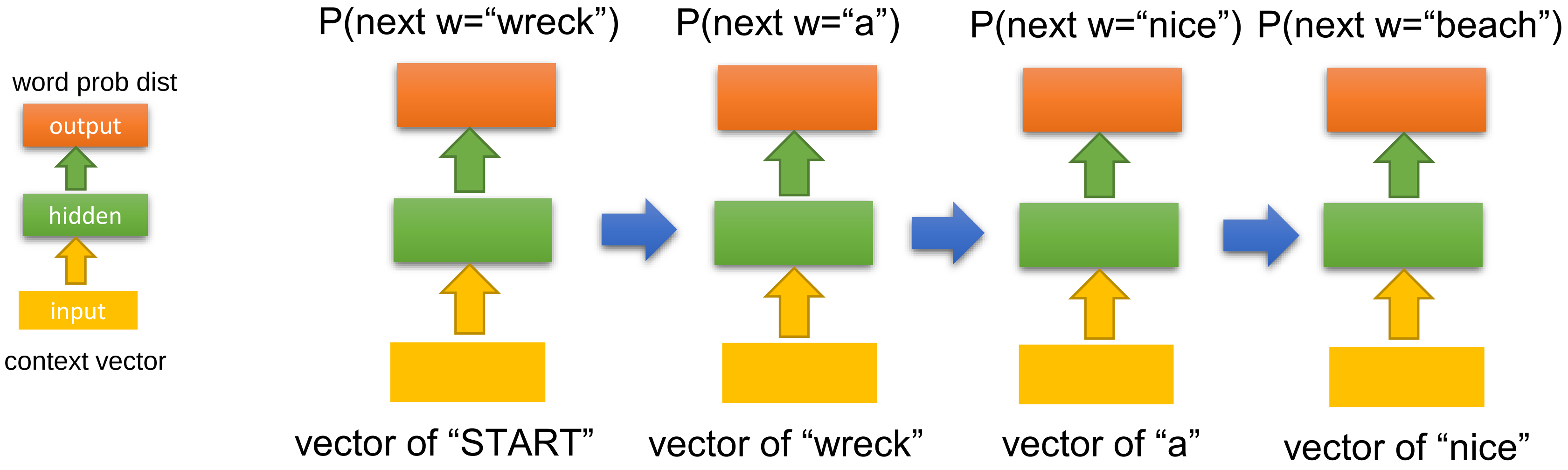
Word Embedding Polysemy Issue

- Words are polysemy
  - ✓ An *apple* a day, keeps the doctor away.
  - ✓ Smartphone companies including *apple*, …
- However, their embeddings are NOT polysemy
- Issue
  - ✓ Multi-senses (polysemy)
  - ✓ Multi-aspects (semantics, syntax)



4 RNNLM

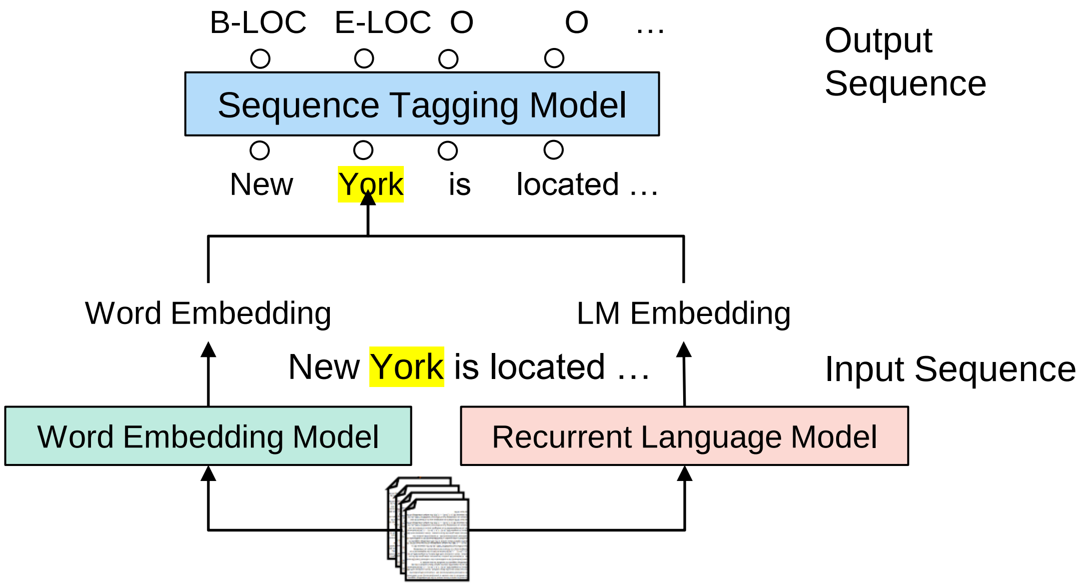
- Idea: condition the neural network on <u>all previous words</u> and <u>tie the weights</u> at each time step



This LM producing **contextual word representations** at each position

TagLM – "Pre-ELMo"

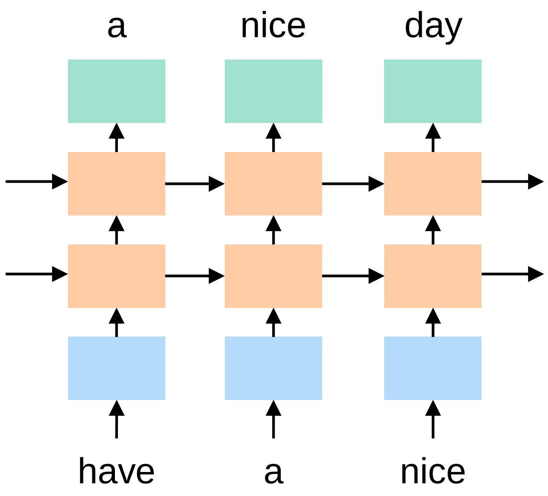
- Idea: train LM on big unannotated data to provide the contextual embeddings for the target task → **self-supervised learning**



Peters et al., "Semi-supervised sequence tagging with bidirectional language models," in *ACL*, 2017.

ELMo: Embeddings from Language Models

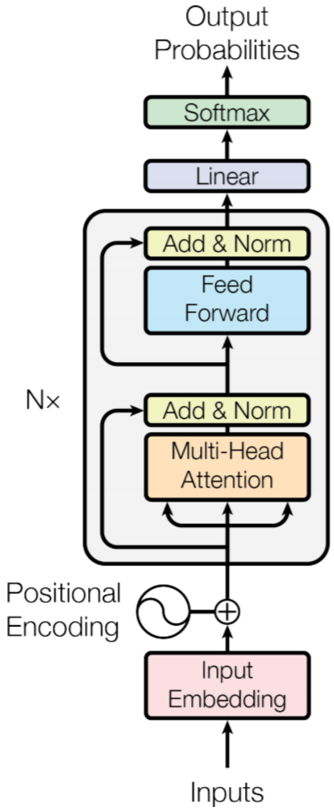
- Idea: contextualized word representations
  - ✓ Learn word vectors using long contexts instead of a context window
  - ✓ Learn a deep LM and use all its layers in prediction



Peters et al., “Deep Contextualized Word Representations”, in *NAACL-HLT*, 2018.

BERT: Bidirectional Encoder Representations from Transformers

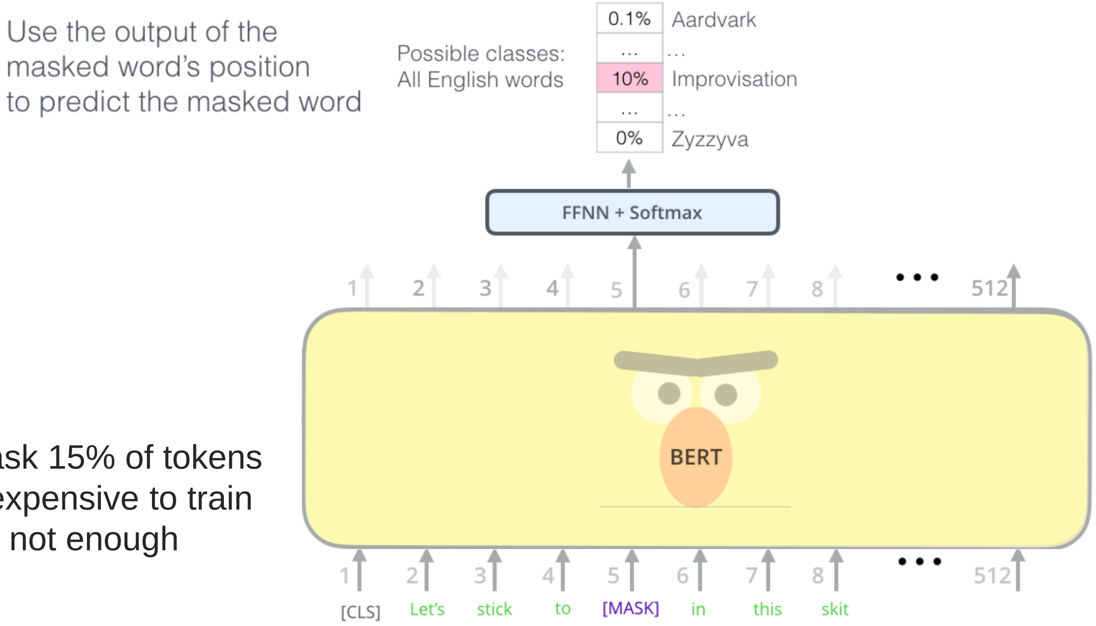
- Idea: contextualized word representations
  - Learn word vectors using long contexts using Transformer instead of LSTM



Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in *NAACL-HLT*, 2019.

BERT #1 – Masked Language Model

- Idea: language understanding is **bidirectional** while LM only uses *left* or *right* context

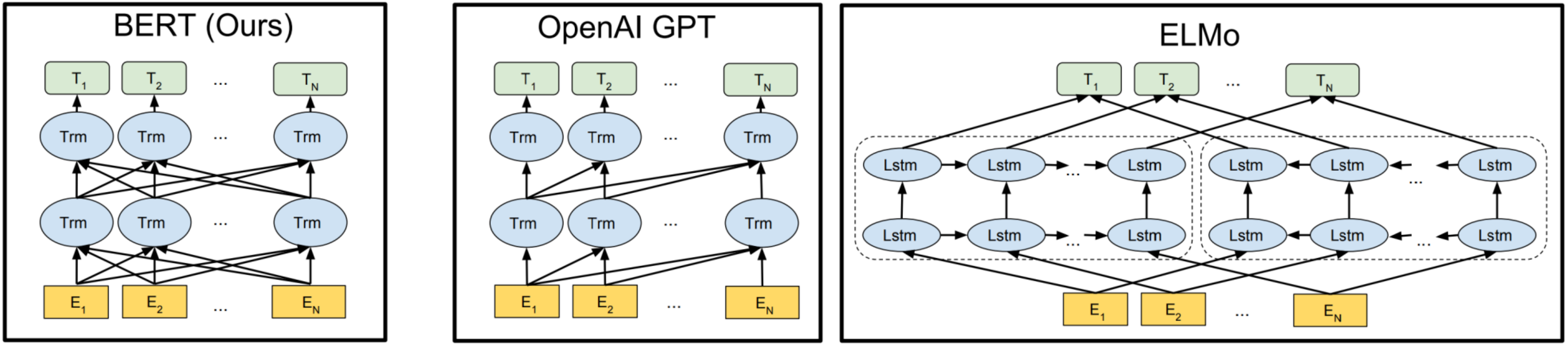


Randomly mask 15% of tokens

- Too little: expensive to train
- Too much: not enough context

http://jalammar.github.io/illustrated-bert/

BERT #1 – Masked Language Model



Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, in *NAACL-HLT*, 2019.

BERT #2 – Next Sentence Prediction

- Idea: modeling ***relationship*** between sentences
  - QA, NLI etc. are based on understanding inter-sentence relationship

Input = `[CLS] the man went to [MASK] store [SEP]`

`he bought a gallon [MASK] milk [SEP]`

Label = `IsNext`

Input = `[CLS] the man [MASK] to the store [SEP]`

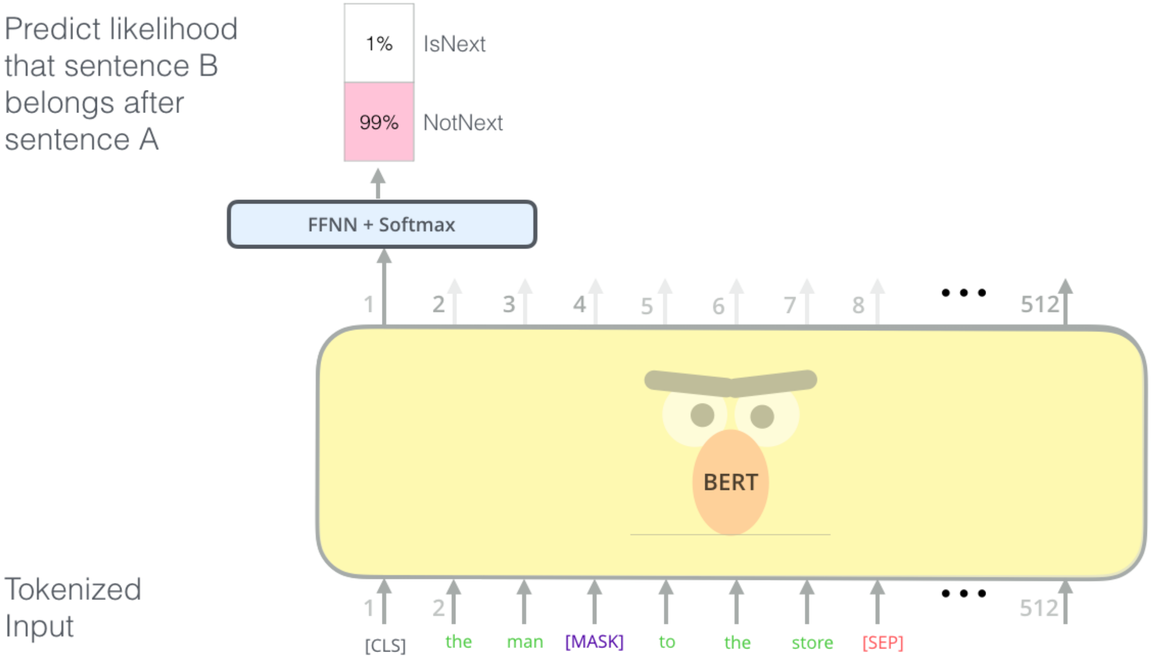
`penguin [MASK] are flight ##less birds [SEP]`

Label = `NotNext`

Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in *NAACL-HLT*, 2019.

BERT #2 – Next Sentence Prediction

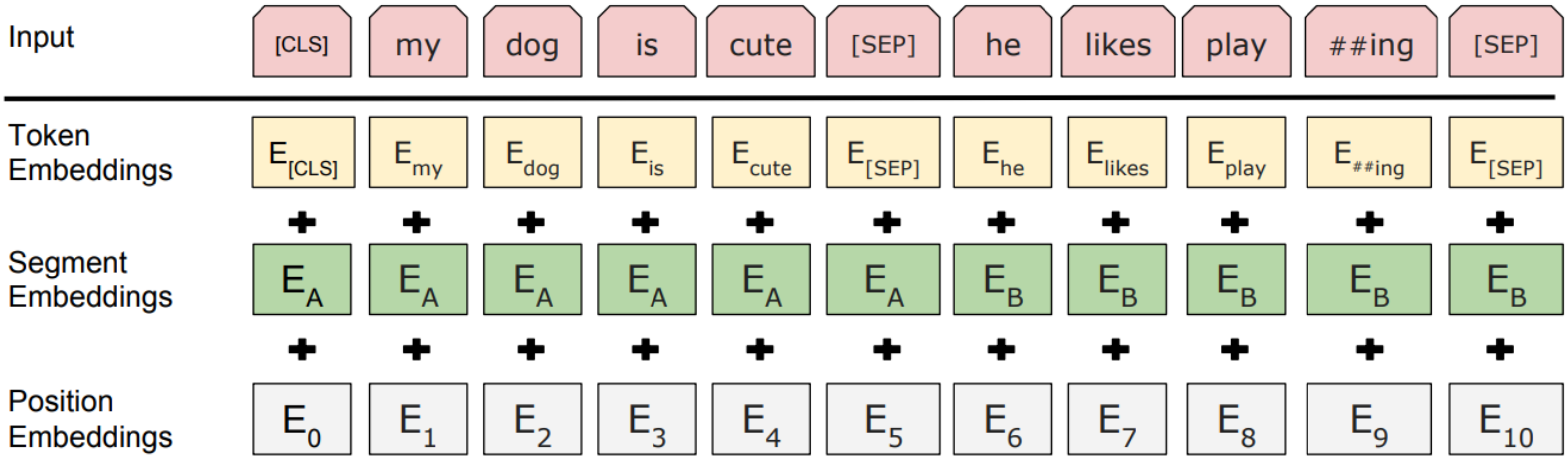
- Idea: modeling relationship between sentences



http://jalammar.github.io/illustrated-bert/

BERT – Input Representation

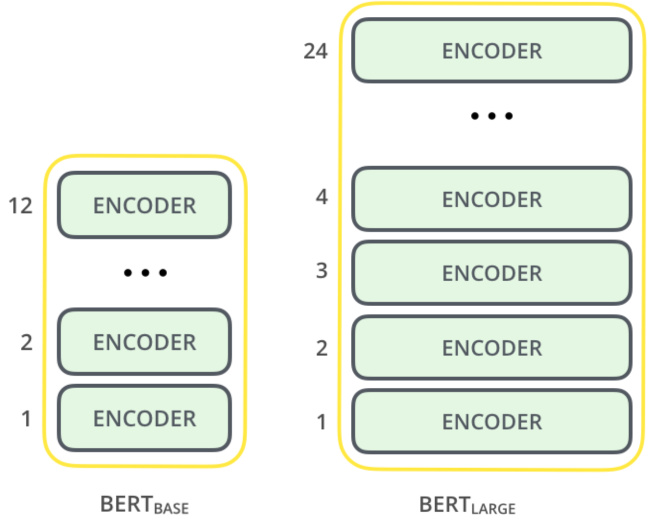
- Input embeddings contain
  - Word-level token embeddings
  - Sentence-level segment embeddings
  - Position embeddings

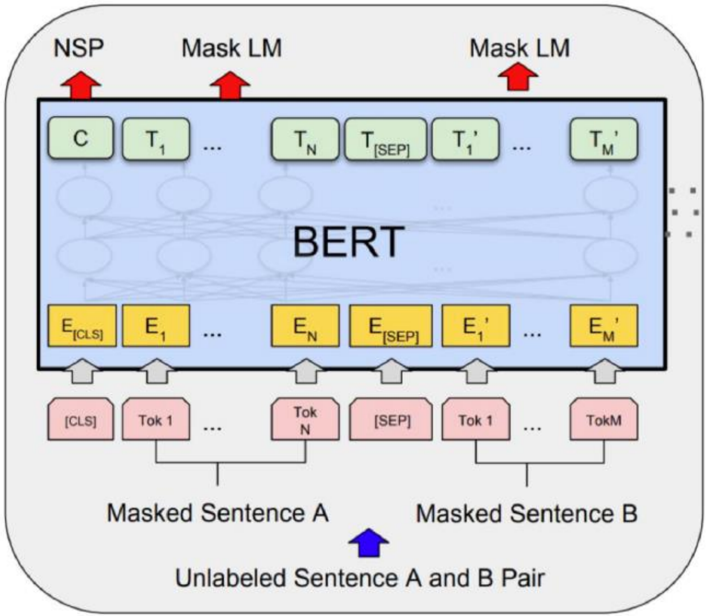


Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in *NAACL-HLT*, 2019.

BERT Training

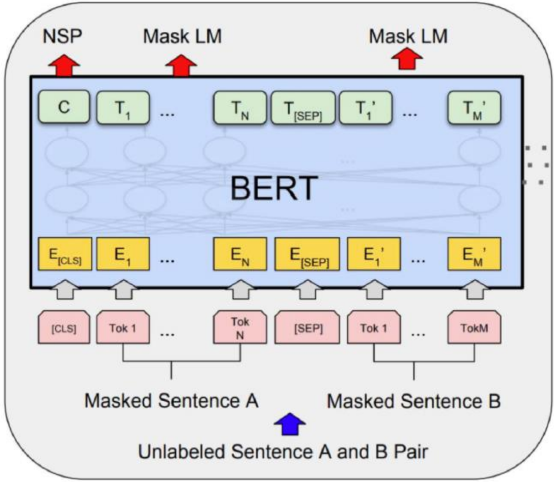
- Training data: Wikipedia + BookCorpus
- 2 BERT models
  - BERT-Base: 12-layer, 768-hidden, 12-head
  - BERT-Large: 24-layer, 1024-hidden, 16-head

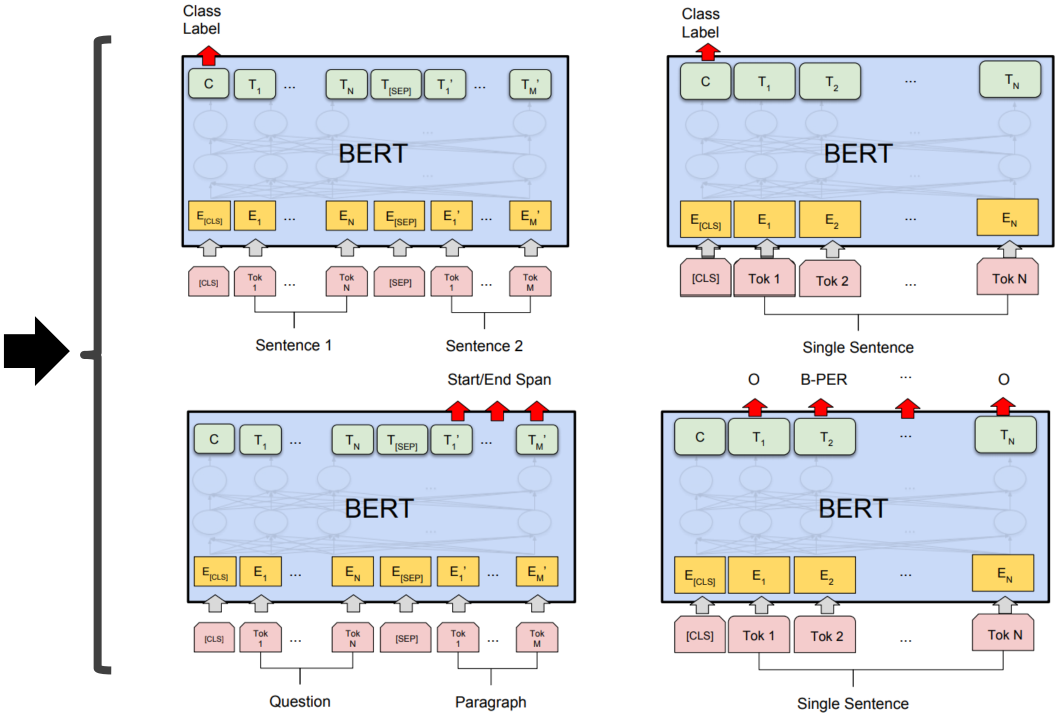




BERT Fine-Tuning for Understanding Tasks

- Idea: simply learn a classifier/tagger built on the top layer for each target task



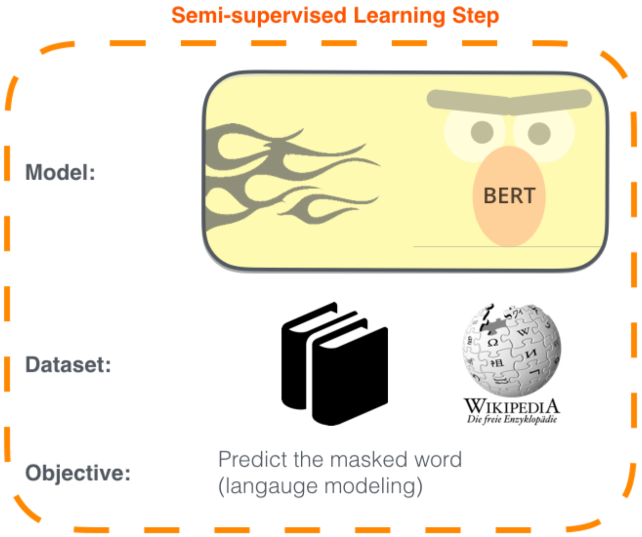


Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in *NAACL-HLT*, 2019.

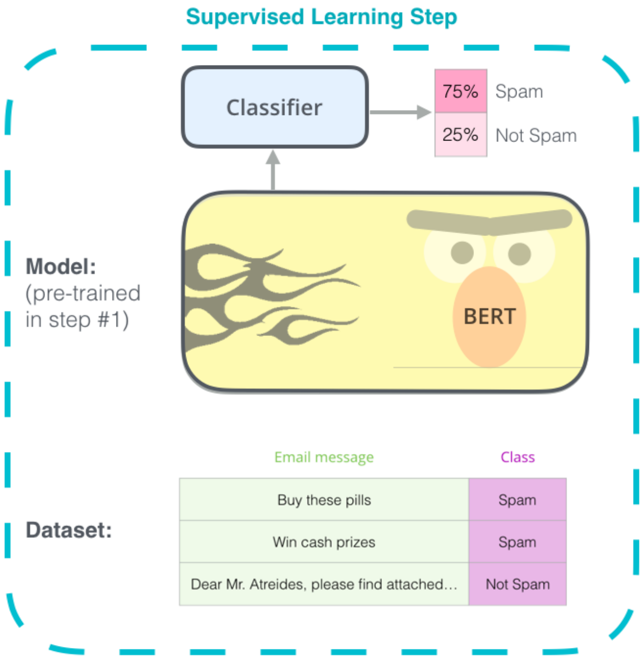
BERT Overview

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

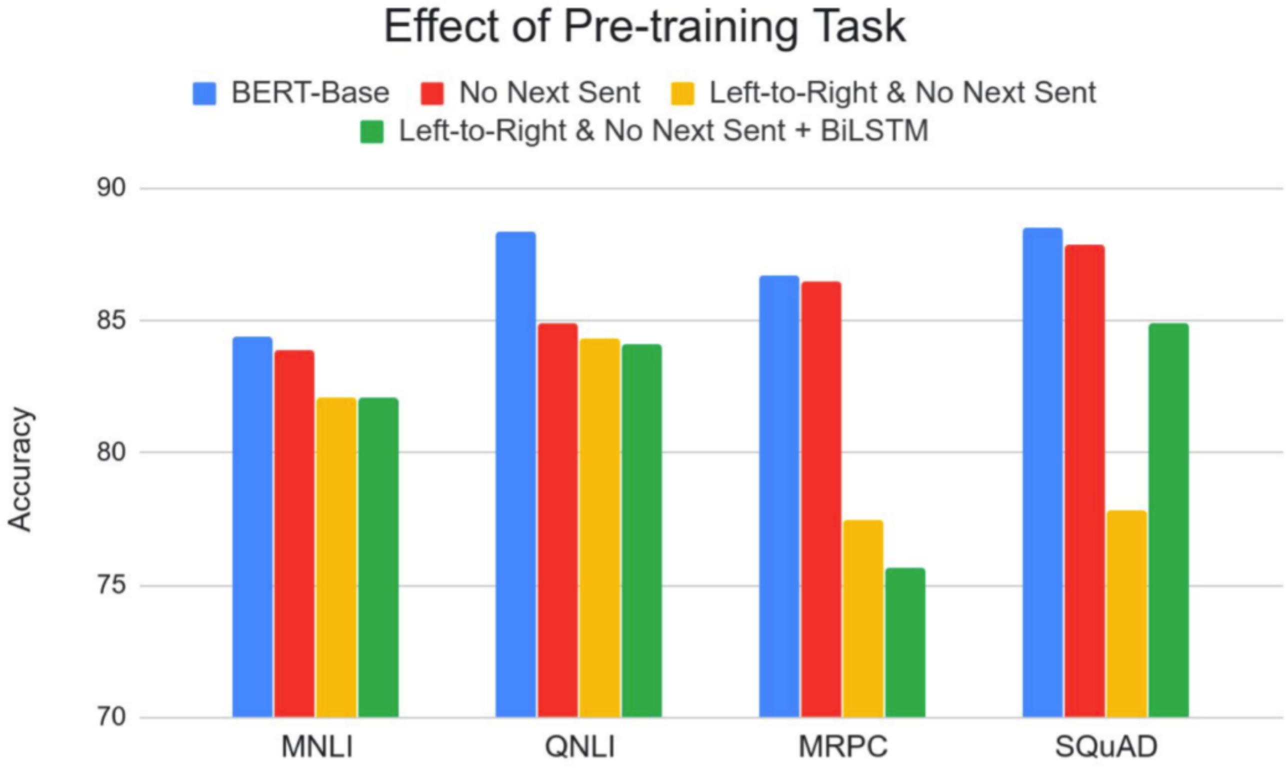


2 - Supervised training on a specific task with a labeled dataset.



http://jalammar.github.io/illustrated-bert/

BERT Fine-Tuning Results



Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in *NAACL-HLT*, 2019.

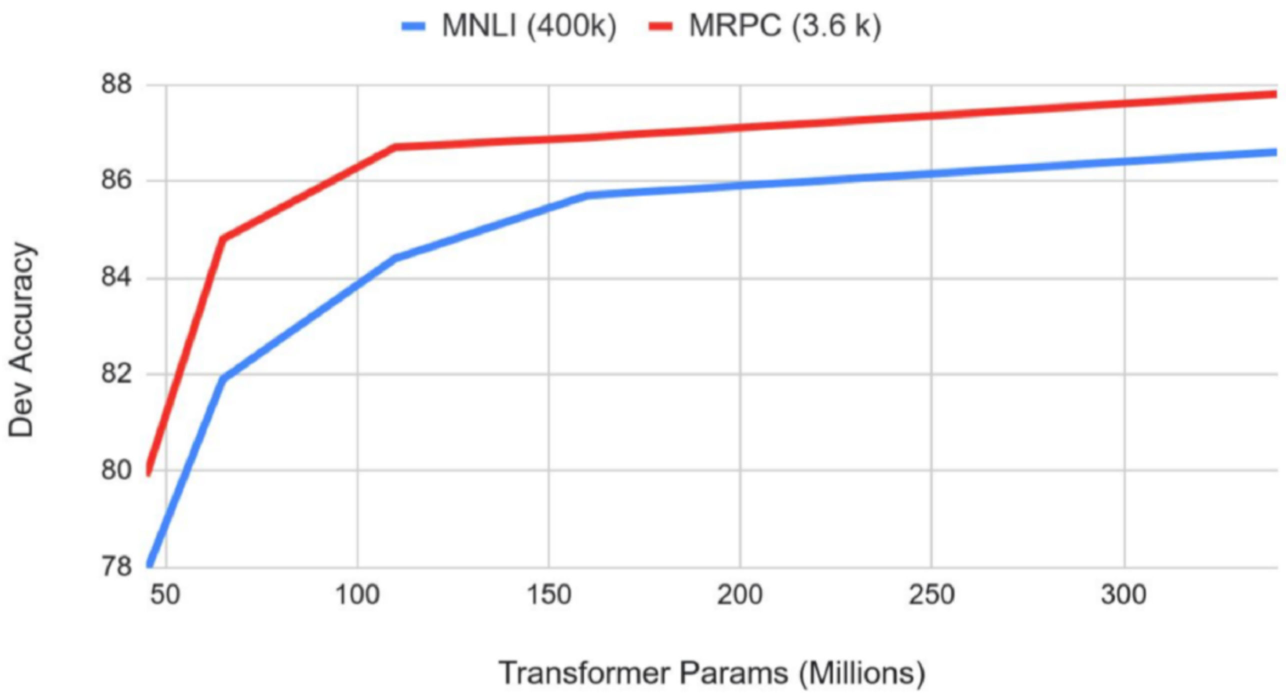
BERT Results on NER

Model	Description	CONLL 2003 F1
TagLM (Peters+, 2017)	LSTM BiLM in BLSTM Tagger	91.93
ELMo (Peters+, 2018)	ELMo in BLSTM	92.22
BERT-Base (Devlin+, 2019)	Transformer LM + fine-tune	<u>92.4</u>
CVT Clark	Cross-view training + multitask learn	92.61
BERT-Large (Devlin+, 2019)	Transformer LM + fine-tune	<u>92.8</u>
Flair	Character-level language model	93.09

Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, in *NAACL-HLT*, 2019.

BERT Results with Different Model Sizes

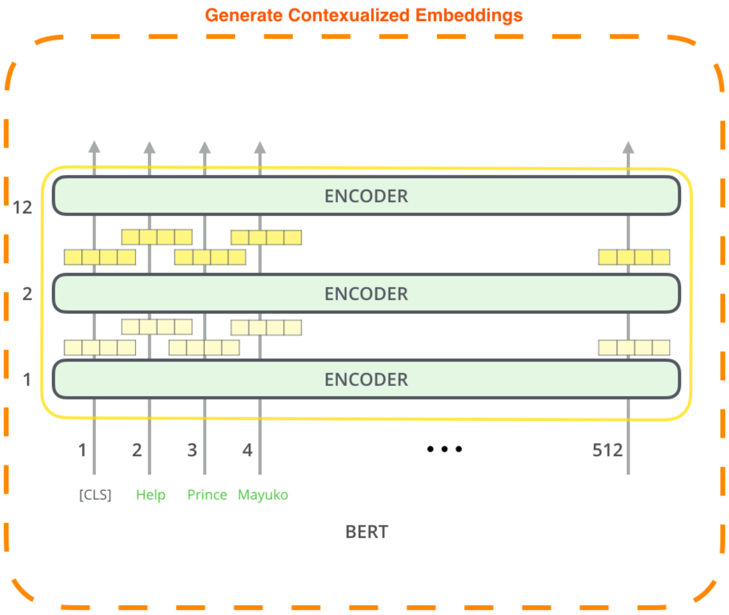
- Improving performance by increasing model size



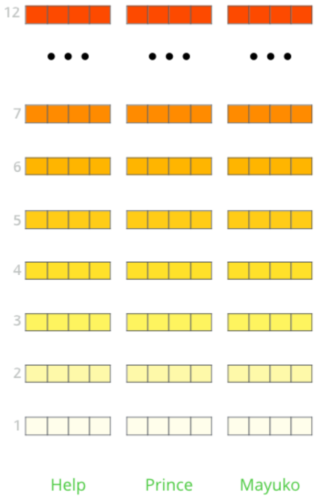
Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, in *NAACL-HLT*, 2019.

BERT for Contextual Embeddings

- Idea: use pre-trained BERT to get contextualized word embeddings and feed them into the task-specific models



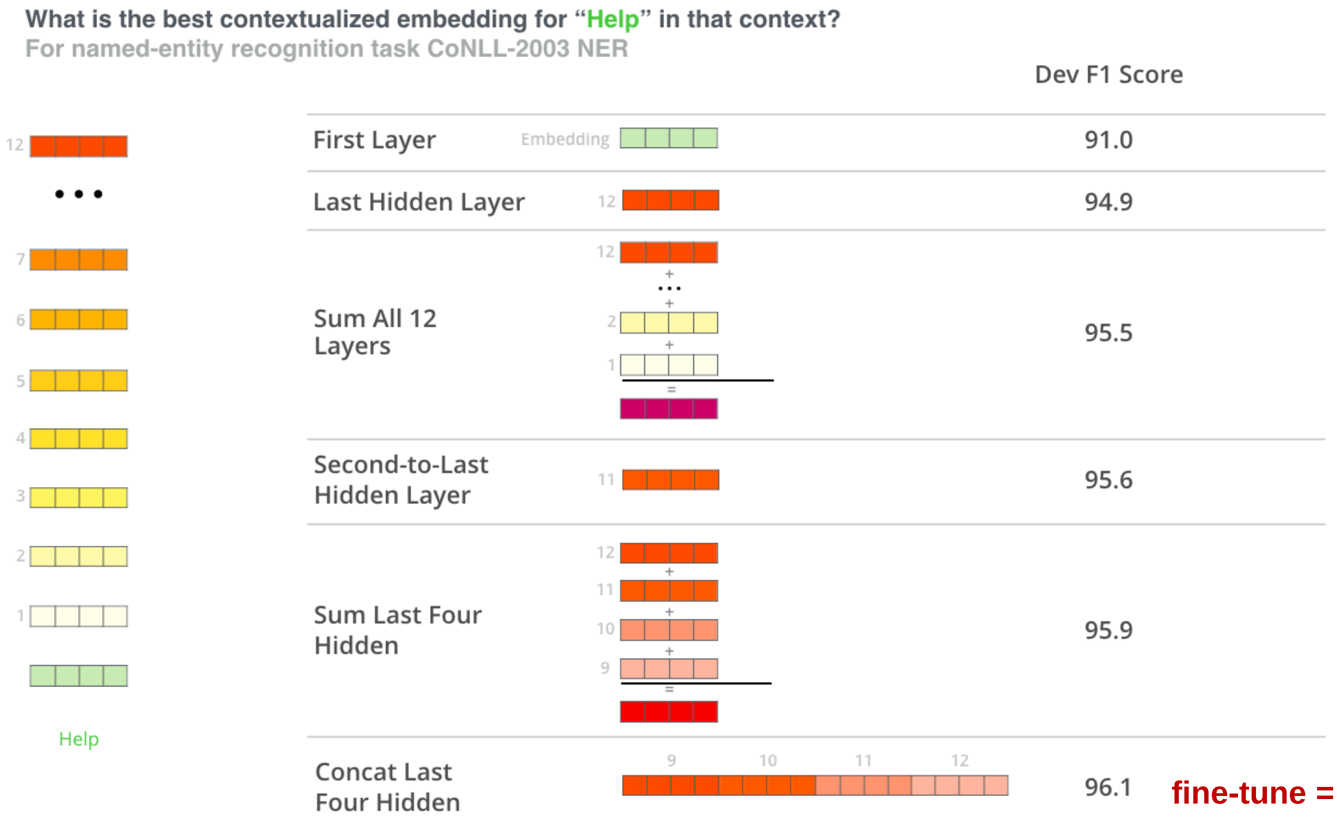
The output of each encoder layer along each token's path can be used as a feature representing that token.



http://jalammar.github.io/illustrated-bert/

But which one should we use?

BERT Contextual Embeddings Results on NER

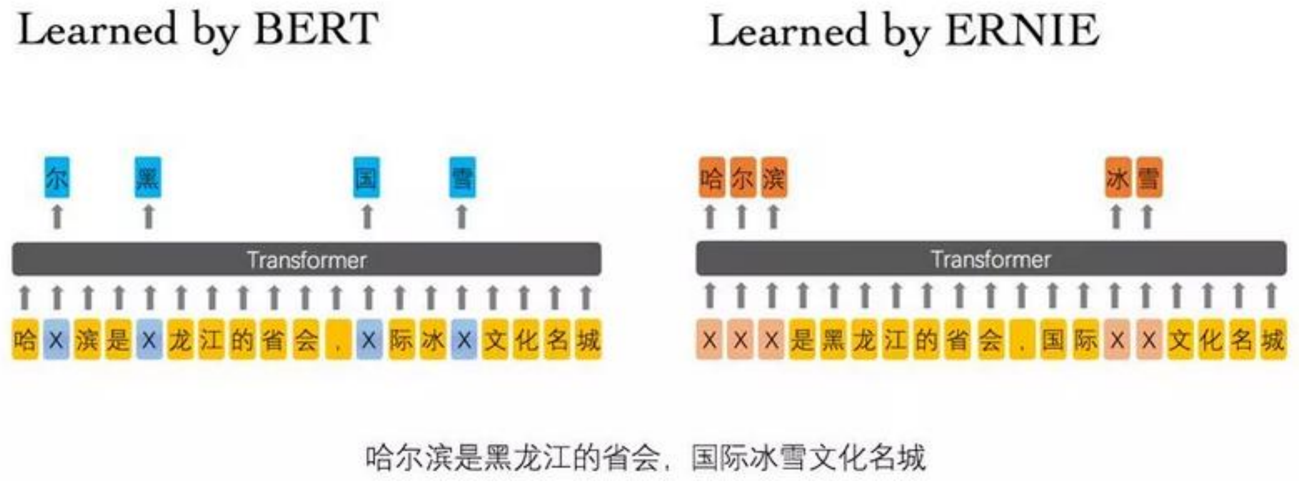


fine-tune = 96.4

http://jalammar.github.io/illustrated-bert/

ERNIE: **E**nhanced **R**epresentation through k**N**owledge **I**nt**E**gration

- BERT models local cooccurrence between tokens, while characters are modeled independently
  - 哈(ha), 爾(er), 濱(bin) instead 哈爾濱(Harbin)
- ERNIE incorporates knowledge by masking semantic units/entities



Concluding Remarks

- Contextualized embeddings learned from masked LM via Transformers provide informative cues for **transfer learning**
- BERT – a general approach for learning contextual representations from Transformers and benefiting language understanding
  - ✓ Pre-trained BERT: https://github.com/google-research/bert https://github.com/huggingface/transformers